Using data mining techniques to predict hospitalization of hemodialysis patients

Jinn-Yi Yeh a,⁎, Tai-Hsi Wu b, Chuan-Wei Tsao a

a Department of Management Information Systems, National Chiayi University, Taiwan
b Department of Business Administration, National Taipei University, Taiwan

ARTICLE INFO

Article history:
Received 20 February 2009
Received in revised form 22 October 2010
Accepted 1 November 2010
Available online 6 November 2010

Keywords:
Hemodialysis
Temporal abstract
Data mining
Healthcare quality

ABSTRACT

Hemodialysis patients might suffer from unhealthy care behaviors or long-term dialysis treatments and need to be hospitalized. If the hospitalization rate of a hemodialysis center is high, its service quality will be low. Therefore, decreasing hospitalization rate is a crucial problem for health care centers. This study combines temporal abstraction with data mining techniques for analyzing dialysis patients’ biochemical data to develop a decision support system. The mined temporal patterns are helpful for clinicians to predict hospitalization of hemodialysis patients and to suggest immediate treatments to avoid hospitalization.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

End stage renal disease (ESRD), commonly known as uremia, is a severe chronic state corresponding to the final stage of kidney failure. In ESRD, kidneys are not able to purify blood from metabolites or to exclude water from the body. Without medical intervention, ESRD patients may die or remain in intensive care unit (ICU) for a long time. These patients require either a kidney transplant or blood-filtering dialysis treatment. The former treatment is difficult to obtain because of a long waiting list and certain patients, such as the elderly, cannot undergo a transplant. The latter includes two main categories, hemodialysis (HD) and peritoneal dialysis (PD). In HD, the blood passes through an extra-corporal circuit where metabolites (e.g. urea) are eliminated. The acid-based equilibrium is re-established and excess water is removed [3]. PD works on the same principles of solute diffusion and fluid ultra filtration as HD, but the blood is cleaned inside the body rather than through a machine [16]. More than 80% of ESRD patients are currently treated with HD [3]. HD patients typically undergo a dialysis session for 4 h, three times a week. During the long-term dialysis treatment, patients will likely receive hospitalization due to caregiver carelessness or other infections. This has been the main reason for HD patient hospitalization in previous years.

High hospitalization rate for a hospital hemodialysis department (HHD) means low service quality in health care. Therefore, the HHD focuses on reducing hospitalization rate. Preventing hospitalization of HD patients from the perspective of preventive medicine is also very important. This paper develops a decision support system to predict hospitalization of HD patients based on a real dataset collected from a hemodialysis center in Taiwan. The HHD examines HD patients receiving long-term treatment to obtain biochemical data during hemodiaysis sessions, such as hematocrit (Hct), albumin, alkaline-p, cholesterol, triglyceride, blood urea nitrogen (BUN), creatinine, uric acid, Na, etc. [25]. The accumulated data over time contains a set of patient variables that are monitored during each dialysis session. The collected data are sequences of multidimensional time series [3].

For time series data, the temporal abstraction (TA) method proposed by Shahar [22] can be integrated with data mining techniques to support data analysis. For example, Bellazzi et al. [3] successfully applied temporal data mining techniques for assessing the clinical performance of HD services such as preprocessing, data reduction, multi-scale filtering, association rule discovery, etc. They found their approach to be suitable for knowledge discovery in clinical time series. Using an auditing system context for dialysis management helped clinicians improve their understanding of patients’ behavior. Adlassnig et al. [1] proposed and discussed promising research directions in the field of TA and temporal reasoning in medicine. They identified and focused on fuzzy logic, temporal reasoning and data mining, health information systems, and temporal clinical databases and recommended developing decision support systems to properly manage the multifaceted temporal aspects of information and knowledge encountered by physicians in their clinical work. Stacey and McGregor [23] surveyed previous research in developing intelligent clinical data analysis systems that incorporate TA mechanisms and present research trends. They suggested the necessity of fusing data mining and TA processes to fully exploit new knowledge from stored clinical data through data mining and apply it to data abstraction.
For TA rule mining, Sacchi et al. [21] proposed a new kind of TA rule and related algorithms for the extraction of temporal relationships between complex patterns defined over time series. Their approaches could be used in a variety of application domains, and they were already tested on two different biomedical problems. Concaro et al. [5] developed a general methodology for the mining of TA rules on sequences of hybrid events for Diabetes Mellitus. The method was capable to characterize subgroups of subjects, highlighting interesting frequent temporal associations between diagnostic or therapeutic patterns and patterns related to the patients’ clinical condition. They concluded that the approach could find a practice for the evaluation of the pertinence of the care delivery flow for specific pathologies.

Based on the literature review, this study integrates TA with data mining techniques for analyzing biochemical data of HD patients to discover temporal patterns resulting in hospitalization. This work develops a decision support system to provide clinicians with association rules and the probability of HD patients’ hospitalization for implementing preventive medicine to decrease hospitalization incidence. This system will hopefully help to understand patients’ changing biomedical data that leads to hospitalization and to improve service quality of the hemodialysis center. The remainder of this paper is arranged as follows: The Materials and methods Section describes hemodialysis and temporal abstraction, the Development of decision support system Section demonstrates the development of the decision support system used in this paper, the Computational results Section illustrates the experimental results using the combined approach for hemodialysis patients’ data analysis, and the last section give the conclusions.

2. Materials and methods

2.1. Hemodialysis

HD for ESRD patients is typically performed in a clinic setting. The diffusion process exchanges solutes and metabolites across a semi-permeable membrane, separating the blood and dialysate. Water is removed from the body using a negative pressure gradient in a process called ultra-filtration. After transit through the dialyzer, the clean, filtered blood is returned to the body. Typically, HD is performed three times a week for about four hours each session. The cost of providing care for HD patients is high. Finding ways to improve patient outcomes and reduce dialysis cost is important. Kusiak et al. [16] demonstrated that data mining, data transformation, data partitioning, and decision-making algorithms are useful for predicting dialysis patient survival. They applied a rough set theory and decision-tree algorithms to analyze biochemical data of HD patients. Sixteen classifiers were produced by these two methods to make predictions. A simple voting scheme was used, with each classifier having one vote. The decision outcome with the maximum number of votes resulted in the predicted outcome. These rules were used to predict the survival of new unseen patients. The results provide a base for analyzing HD patients’ data in Taiwan.

The required HD service for ESRD patients has dramatically increased each year in Taiwan. Due to the special characteristics of these patients, both hospitalization times and hospitalization costs are higher than for other patients. Preventive medicine could decrease these costs by analyzing HD lab data to find possible factors. The Bureau of National Health Insurance in Taiwan has also put forward a professional health care quality index according to various overall quality schemes for long-term monitoring. The main items for assessing hemodialysis center quality include: hospitalization rate, serum albumin, clearance rate of urea nitrogen (Kt/V), hematocrit, death rate, sinus reconstruction rate, and weaning rate. Hong [11] used multiple minimum association rules to discover hospitalization cause and effect factors by analyzing HD lab data. His approach increased the accuracy of health care results and shortened the analyzing period. However, he did not consider time series data recorded during HD sessions.

2.2. Temporal abstraction

Temporal abstraction (TA) is an artificial intelligence technique, which integrates domain knowledge into the data analysis process. TA outlines the evolutionary process of temporal data through a qualitative presentation mode, such as level shifts, periods of stability and trends. Shahar [22] defined TA as a program given a set of time series data including variables, external events, and abstract. The generated abstract description represents previous and current states and data trends. The TA program converts patients’ data from a low-level quantitative format to a high-level qualitative description. This presentation format is close to the clinician’s specialty vocabulary [23]. Clinicians and domain experts typically work together to discuss rules and knowledge based on TA. These rules and knowledge are very important for generating significant and data-dependent TA and to determine whether these abstractions can be explained correctly to work out correct diagnoses.

Generally, TA can be obtained from both basic TA and complex TA, described respectively as follows. Basic TA is typically indicated by combining state and trend from time series data of existing episodes. The state can be classified as low, normal, and high values. The trend can be classified as increased, decreased, and stable patterns [22]. An episode refers to the data in a time interval [3]. Complex TA describes the temporal relation between basic TAs and complex TAs. Typical complex TAs typically use the temporal operator proposed by [2] to concatenate basic TA (see Fig. 1). The most used operator is Meet, referring to the successively presented precedence order between basic TAs [23]. For example, if a rule states that two basic TAs, A-high and B-high, correspond to result C, and A occurs earlier than B, it can be indicated by A-high Meet B-high Then C.

Initially, TA is applied to data monitoring of patients in intensive care units (ICU) based on the intelligent data analysis (IDA) system which detects abnormal phenomena of patients’ temporal data. This provides clinicians with relevant temporal information of patients for subsequent treatments [17]. Fig. 2 shows the IDA system schema, containing several different missions: data validation, data representation, data interpretation, and control tasks. TA is applicable to data representation and the interpretation phase. When TA is applied to the data analysis system during the reasoning process, interpretation or reasoning can be carried out by comparing predefined patterns defined by clinicians or derived from machine learning techniques. Finally, the control tasks present proper treatments according to abnormal phenomena of the data.

<table>
<thead>
<tr>
<th>Relation</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>A operator(PRECEDE) B</td>
<td>Overlap : aaaaa bbbbbb</td>
</tr>
<tr>
<td></td>
<td>Meet: aaaa bbbbbb</td>
</tr>
<tr>
<td></td>
<td>Before : aaaaa bbbbbb</td>
</tr>
<tr>
<td></td>
<td>Equal : aaaaa bbbbbb</td>
</tr>
</tbody>
</table>

Fig. 1. Temporal operator [3].
TA application is also related to medical data analysis, and often integrated with data mining techniques and artificial intelligence approaches. In the study of [3], blood pressure and ferroheme of hemodialysis patients were derived from dialyzers in three weekly treatments. They were processed by discrete wavelet transformation and association rules mining to obtain insufficient factors of hemodialysis service quality. The mined rules provided a reference for clinicians. The results show that an auditing system context for dialysis management helped clinicians improve their understanding of patients' behavior. Differing from [3], the hepatitis data used in [10] was collected at irregular intervals, therefore the TA time intervals were not the same. The study also used a “state change” temporal operator to describe temporal abstraction, and formula counting significance to prune the rules. The experimental results show that the generating hepatitis temporal pattern improved effectively.

2.3. Decision tree

A decision tree (DT) is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node represents hospitalization or non-hospitalization. The top-most node in a tree is the root node.
test on an attribute is associated with a splitting criterion chosen to split the data sets into subsets with better class separability, thus minimizing misclassification error. Once the tree is built from the training data, it is then heuristically pruned to avoid over-fitting of data, which tends to introduce classification error on the test data [9,14].

For DT pruning, C4.5 follows the postpruning approach which removes branches from a fully grown tree. For each nonleaf node in the tree, the pruning algorithm estimates the expected error rate that would occur if the subtree at that node were pruned. Then the expected error rate occurring if the node were not pruned is estimated using the error rates for each branch, combined by weighting according to the proportion of observations along each branch. If pruning the node leads to a greater expected error rate, then the subtree is kept. Otherwise, it is pruned [9].

2.4. Mining association rules with multiple minimum supports

Mining association rules with multiple minimum supports is an important generalization of the association-rule-mining problem, recently proposed by [19]. Instead of setting a single minimum support threshold for all items, users are allowed to specify multiple minimum supports to reflect the nature of items, and an Apriori-based algorithm is developed to mine all frequent item sets. Thus, different rules may need to satisfy different minimum supports depending on what items are in the rules. This new model enables users to produce rare item rules without causing frequent items to generate too many meaningless rules [12].

3. Development of decision support system

The schema system of this study was mainly developed from IDA, shown in Fig. 3. In the knowledge creation phase, patient’s time series biochemical data is validated by knowledge engineers and relevant clinicians to remove outliers and to handle missing data. The preprocessed data is changed to a TA format based on TA rules specified by the domain expert. This process integrates the knowledge of specialized fields into the analysis program. The transformed TA data is afterwards fetched into the data mining program to find out rules for predicting hemodialysis patients’ hospitalization. The mined patterns are provided to clinicians for further processing, including rules evaluation and irrational rules deletion. The remaining rules are stored in a database, to provide a decision support system for clinicians to practically judge whether patients will need hospitalization or not.

3.1. Data collection

This study was conducted in a large nationwide hemodialysis center in Taiwan. A total of 8223 chronic hemodialysis patients (samples) who received regular hemodialysis for more than six consecutive months comprised the analysis subject. Data collection was based on known and unknown biochemical testing items (parameters) of effective dialysis treatment. The data mining view considers that every biochemical testing item represents a feature. Due to different biochemical testing items having their own clinical significance, we needed to discuss with a clinician to sort out the applications of various biochemical testing items. The clinical significance of some biochemical testing items in the data also appears frequently. After making enquiry, we selected the feature and avoided redundant features influencing mining accuracy and effectiveness. For example, protein and albumin represent patients’ nutritional status. Clinicians suggest selecting any of them as the feature.

Features used in this study were collected from 2005–2007 quarterly reports including gender, age, hospitalization, admission date, and monthly routine biochemical test items, such as White Blood Cell (WBC), Red Blood Cell (RBC), Hemoglobin (HBC), Hematocrit

![Fig. 4. Data preprocessing flow chart.](image-url)
(HCT), Mean Corpuscular Volume (MCV), Platelet, Albumin, Alanine aminotransferase (GPT), Alkaline Phosphatase (Alkaline-P), Cholesterol, Triglyceride, Glucose (AC), Creatinine, Uric Acid, Sodium (Na), Potassium (K), Calcium (Ca), Phosphorus (P), Urea Reduction Rate (URR), kt/v, Serum iron (Fe), Ferritin, and Ca×P where URR was calculated from the difference in blood urea concentration before and after dialysis divided by pre-dialysis blood urea concentration. The KT/V (Daugirdes) was calculated using the URR and adjusted for fluid shifts that occur during dialysis. The product of calcium and phosphorous (Ca×P) has special meaning. If serum inorganic phosphorus is greater than 5.5 and Ca×P is greater than 55, death rate increases accordingly (calcification of main coronary artery and cardiac muscle result in coronary heart disease and heart failure). Therefore the value of Ca×P needs additional calculations added to the system.

### 3.2. Data preprocessing

Due to heterogeneity and specialty problems of medical data, we analyzed the biochemical tests of original HD patients during the preprocessing phase. We also found it necessary to cooperate with clinicians in this step to find out outliers and missing values. The process is shown in Fig. 4.

#### Table 1

<table>
<thead>
<tr>
<th>No.</th>
<th>Item</th>
<th>Normal value Unit</th>
<th>XH</th>
<th>H</th>
<th>N/H</th>
<th>N/L</th>
<th>L</th>
<th>XL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>WBC</td>
<td>5–10×1000/ul</td>
<td>30</td>
<td>15</td>
<td>10</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>RBC</td>
<td>&gt;2.5×10^6/ul</td>
<td>6×.5</td>
<td>6</td>
<td>5.4</td>
<td>3.8</td>
<td>3</td>
<td>2.5</td>
</tr>
<tr>
<td>3</td>
<td>HBC</td>
<td>12.3–18.3 g/dl</td>
<td>20</td>
<td>18</td>
<td>15</td>
<td>9</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>HCT</td>
<td>39–53%</td>
<td>50</td>
<td>45</td>
<td>40</td>
<td>28</td>
<td>25</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>MCV</td>
<td>80–100 fl</td>
<td>110</td>
<td>100</td>
<td>95</td>
<td>80</td>
<td>70</td>
<td>60</td>
</tr>
<tr>
<td>6</td>
<td>Platelet</td>
<td>120–320×1000/ul</td>
<td>450</td>
<td>400</td>
<td>360</td>
<td>100</td>
<td>50</td>
<td>30</td>
</tr>
<tr>
<td>7</td>
<td>Albumin</td>
<td>3.7–5.1 gm/dl</td>
<td>5.5</td>
<td>5</td>
<td>4.8</td>
<td>3.4</td>
<td>3</td>
<td>2.5</td>
</tr>
<tr>
<td>8</td>
<td>GPT</td>
<td>0–36 mg/dl</td>
<td>500</td>
<td>100</td>
<td>45</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>Alkaline-P</td>
<td>37–95 IU/l</td>
<td>200</td>
<td>150</td>
<td>129</td>
<td>40</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td>10</td>
<td>Cholesterol</td>
<td>&lt;200 mg/dl</td>
<td>260</td>
<td>240</td>
<td>200</td>
<td>130</td>
<td>120</td>
<td>100</td>
</tr>
<tr>
<td>11</td>
<td>Triglyceride</td>
<td>30–160 mg/dl</td>
<td>500</td>
<td>300</td>
<td>180</td>
<td>120</td>
<td>100</td>
<td>50</td>
</tr>
<tr>
<td>12</td>
<td>Glucose(AC)</td>
<td>80–120 mg/dl</td>
<td>500</td>
<td>300</td>
<td>110</td>
<td>70</td>
<td>60</td>
<td>50</td>
</tr>
<tr>
<td>13</td>
<td>Creatinine</td>
<td>0.7–2.0 mg/dl</td>
<td>20</td>
<td>15</td>
<td>10</td>
<td>8</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>14</td>
<td>Uric Acid(Male)</td>
<td>2.7–8.5 mg/dl</td>
<td>10</td>
<td>8</td>
<td>7.6</td>
<td>3.4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>Uric Acid(Female)</td>
<td>3.2–8.2 mg/dl</td>
<td>9</td>
<td>7</td>
<td>6</td>
<td>2.8</td>
<td>2.5</td>
<td>2</td>
</tr>
<tr>
<td>16</td>
<td>Na</td>
<td>134–138.5 meq/l</td>
<td>160</td>
<td>155</td>
<td>150</td>
<td>130</td>
<td>120</td>
<td>110</td>
</tr>
<tr>
<td>17</td>
<td>K</td>
<td>3.4–4.8 m/l</td>
<td>7.5</td>
<td>6.5</td>
<td>6</td>
<td>3</td>
<td>2.5</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>Ca</td>
<td>8.4–10.2 mg/dl</td>
<td>12.5</td>
<td>10</td>
<td>9.5</td>
<td>8.4</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>19</td>
<td>P</td>
<td>2.8–112 mg/dl</td>
<td>7</td>
<td>6</td>
<td>5.5</td>
<td>3.5</td>
<td>3</td>
<td>2.5</td>
</tr>
<tr>
<td>20</td>
<td>URR</td>
<td>0.65</td>
<td>0.8</td>
<td>0.75</td>
<td>0.7</td>
<td>0.6</td>
<td>0.55</td>
<td>0.5</td>
</tr>
<tr>
<td>21</td>
<td>kt/v(Daugirdes)</td>
<td>1.3</td>
<td>1.6</td>
<td>1.45</td>
<td>1.35</td>
<td>1.2</td>
<td>1</td>
<td>0.8</td>
</tr>
<tr>
<td>22</td>
<td>Fe</td>
<td>50–190 mg/dl</td>
<td>200</td>
<td>160</td>
<td>135</td>
<td>80</td>
<td>50</td>
<td>30</td>
</tr>
<tr>
<td>23</td>
<td>Ferritin</td>
<td>100–300 ng/ml</td>
<td>1000</td>
<td>800</td>
<td>650</td>
<td>300</td>
<td>150</td>
<td>50</td>
</tr>
<tr>
<td>24</td>
<td>Transferrin saturation</td>
<td>335</td>
<td>60</td>
<td>50</td>
<td>40</td>
<td>30</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>25</td>
<td>intact_PTH</td>
<td>130–195 pg/ml</td>
<td>1000</td>
<td>600</td>
<td>300</td>
<td>60</td>
<td>50</td>
<td>40</td>
</tr>
<tr>
<td>26</td>
<td>Ca×P</td>
<td>50</td>
<td>70</td>
<td>65</td>
<td>55</td>
<td>45</td>
<td>35</td>
<td>30</td>
</tr>
</tbody>
</table>
Since HD patients’ physiology is different from that of a normal person, a minority of patients’ routine biochemical test values are ultra-abnormal. Medical data often support this fact. Furthermore, when clinicians input biochemical test values into computers, it is hard to avoid typos, resulting in serious data analysis errors. Therefore, it is also necessary to contact clinicians about the standard value scope of various testing items in this step. Then outliers can be found based on three times of standard deviation [14]. After deleting spurious data as well as outliers, 6776 samples remained, including 3146 males and 3630 females. The average age was 59.92 and the standard deviation was 13.6. Out of 6776 patients, 1710 (25.24%) were hospitalized during their HD treatment sessions.

Because the DT algorithm does not allow null values, we discarded more than 15 percent of those features with missing data, such as Chloride (Cl), Normalized Protein Catabolic Rate (nPCR), Unsaturated Iron Binding Capacity (UIBC), Total Iron Binding Capacity (TIBC), Tranferrin saturation, Aluminum (Al), Magnesium (Mg), intact-PTH, and Cardiac/thoracic rate. For those features with less than 15 percent missing data, we filled up the mean value of previous and latter values of the patient. After deleting those values containing nulls and missing data, 6284 samples remained.

3.3. Transforming to temporal abstraction

To change the monthly biochemical test data of HD patients into TA, we developed a system to extract professional persons’ HD knowledge. The conversion rule threshold for temporal abstraction inputted by clinicians is visually presented. Clinicians can also indicate the clinical significance of the biochemical testing item in the remarks column. The TA primitives are defined as follows [24]:

1. State primitive: N (normal), L (low), VL (very low), XL (extreme low), H (high), VH (very high), and XH (extra high).
2. Relations: “=” (change state to), “&” (and), “-” (and then), and “/” (X/Y means the majority of points are in state X and the minority of points are in state Y).

Fig. 5 shows the transformation flow. Each biochemical test item should obtain its corresponding normal value in advance to define the threshold values for transforming to TA. Table 1 lists these data. A time interval should also be determined to transform data to TA. Hong [11] set the time interval of basic TA at three months. However, to find out the evolutionary patterns which may cause patients’ hospitalization in six months of HD treatment, our study defined the complex TA as six months for one interval. A long-term HD patient that does not receive hospitalization within six to twelve months after treatment would be regarded as a non-hospitalization data set. A patient that receives hospitalization within six months after treatment would be regarded as a hospitalization data set.

This study transformed patients’ biochemical test value data into basic TA through the basic TA algorithm, and inputted the result into the complex TA algorithm to figure out the complex TA described as follows.

3.3.1. Transforming to basic TA

The time series data of HD patients should be transformed to qualitatively describe with time interval before data mining, to execute temporal mining. Firstly we need to define the basic TA format according to the abstraction rules given by clinicians (as Fig. 6).

For example, we assume that the time series data of biochemical test values of a patient is recorded from January to October, shown in Fig. 7. According to the abstraction intervals of basic TA and complex TA, we can select a set of basic TA and a set of complex TA from the data.

3.3.2. Transforming to complex TA

Composing the basic TA to complex TA can be concatenated by the temporal operator. The temporal operator, > (from one state to another), states transforming and trend evolution. This study used the temporal operator put forward by Ho et al. [10] to concatenate the basic TA. For example, the patient’s long-term representation can be indicated by complex TA "normal (N) > higher than normal (N/H)". The patient’s biochemical test value changes from normal to higher, than to normal state from the first month to the sixth month. This also means the patient’s biochemical test value tends to climb. Therefore, the complex TA formed by using the temporal operator contains the basic state and trend abstraction at the same time. After obtaining the basic TA, we can input all patients’ basic TA data into the complex TA algorithm to transform to temporal abstraction. The complex TA algorithm is as follows:

Complex TA algorithm

Input: Seasonal (S) Biochemical test items’ (Bio) Basic TA (State_TA), current month (Near_HD_date), and hospitalization date (Hospital_date) of each patient (P)

Parameters: Hospitalization (Yes/No)

Output: P.Complex_TA

1 for each P
2 
3 if all Hospitals of P=N then
4 Judge S belongs in which Season according to Near_HD_date
6 P.Complex_TA=Temp_V
7 else
8 Judge S belongs in which Season according to Hospital_date
10 P.Complex_TA=Temp_V
11 end if
12 next
13 next

After executing the complex TA algorithm, Table 2 lists the data format for further analyzing, which will be inputted into the
MSApriori algorithm and the C4.5 decision tree for hospitalization pattern mining.

3.4. Hospitalization rule mining

This study applied association rule mining with multiple minimum supports and the DT to analyze TA data of HD patients, and finds out rules for predicting patients’ hospitalization, to prevent hospitalization of HD patients, described as follows.

3.4.1. Mining association rules with multiple minimum supports

Using Apriori for association rules mining results in a shortage. For example, in the retailing business, if the minimum support is set too high, all the discovered rules are concerned with low-price products, which only contribute a small portion of the profit to the business. On the other hand, if minimum support set too low, many meaningless frequent rules will be generated that will overload the decision maker, making it difficult to understand the rules [7]. Therefore, Liu et al. [19] proposed the MSApriori algorithm using multiple minimum supports for different items based on the Apriori architecture. The result was actually more effective than Apriori because its association rules with some important and infrequent item set were found.

Similar to the retail business, the MSApriori algorithm is suitable for analyzing HD patients’ data because some TA items are closely related to patients’ hospitalization rate such as Kt/V, URR, albumin etc. These items are so important, that the item’s minimum support must be particularly considered by the clinician. Therefore, this study used the MSApriori algorithm to find out the most relevant TA of the testing item causing hospitalization. Minimum item support (MIS) represents the minimum support for each testing item defined as follows:

$$\text{MIS}(i) = \frac{\text{TA}(i) \cup H_{\text{yes}}}{N}$$  \hspace{1cm} (4)

where $i$ represents the testing item, TA($i$) is the TA of each testing item, i.e. $N \rightarrow N/H$, $H_{\text{yes}}$ indicates that the hospitalization type is yes, $N$ is the number of samples, and $\cup$ means the union of sets TA($i$) and $H_{\text{yes}}$, or say both TA($i$) and $H_{\text{yes}}$. If a TA’s MIS is less than the threshold 0.1, then the TA’s MIS is set at 0.1. To evaluate correlativity between the testing item and hospitalization, Conf is defined as follows [3]:

$$\text{Conf}(\text{TA}(i) \rightarrow H_{\text{yes}}) = \frac{P(H_{\text{yes}} \mid \text{TA}(i))}{P(\text{TA}(i))} = \frac{\text{support}(\text{TA}(i) \cup H_{\text{yes}})}{\text{support}(\text{TA}(i))}$$ \hspace{1cm} (5)

where TA($i$) $\rightarrow$ H$_{\text{yes}}$ is a rule and support(TA($i$)) is the percentage of samples that contain TA($i$). This study used the cosine measure proposed by Han and Kamber [9] to delete rules. The cosine measure of TA($i$) and H$_{\text{yes}}$ is defined as

$$\text{cosine}(\text{TA}(i), H_{\text{yes}}) = \frac{P(\text{TA}(i) \cup H_{\text{yes}})}{\sqrt{P(\text{TA}(i)) \times P(H_{\text{yes}})}} \hspace{1cm} (6)$$

Let $I_k$ denote the set of large k-item sets. Each item set $c$ is of the following form, $\langle c[1], c[2], ..., c[k] \rangle$, which consists of items, $c[1], c[2], ..., c[k]$, where $\text{MIS}(c[1]) \leq \text{MIS}(c[2]) \leq ... \leq \text{MIS}(c[k])$. The MSApriori algorithm is listed as follows [11]:

**MSApriori algorithm**

1. $M=\text{sort}(I, \text{MS})$ /* according to MIS($i$)’s stored in MS */
2. $T=\text{int-pass}(N, \text{T})$ /* make the first pass over T */
3. $L_0=\{f\mid f \in F, f.\text{count} \geq \text{MIS}(f)\}$
4. for ($k=1; L_{k-1} \neq \emptyset; k++$) do
5. if $k=2$ then $C_2=\text{level2-candidate-gen}(F)$
6. else $C_k=\text{candidate-gen}(L_{k-1})$ end
7. for each transaction $t \in T$ do
8. $C_t=\text{subset}(C_k, t)$
9. for each candidate $c \in C_t$ do $c.\text{count}++$
10. end
11. $L_k=\{c \in C_k \mid c.\text{count} \geq \text{MIS}(c[1])\}$
12. end
13. Answer=$\bigcup L_k$

Algorithm level2-candidate-gen

**Input:** $F$

**Output:** a superset of the set of all large 2-itemsets

1. for each item $f$ in $F$ in the same order do
2. if $f.\text{count} \geq \text{MIS}(f)$ then
3. for each item $h$ in $F$ that is after $f$ do
4. if $h.\text{count} \geq \text{MIS}(f)$ then
5. insert $<f, h>$ into $C_2$

3.4.2. Decision tree

The DT result determines which items can be used as the feature prediction model for predicting classification accuracy. We used InfoGainAttributeEval as the index for calculating the significance level of each biochemical item, and used the Ranker method to rank all biochemical items according to their significance levels [26]. The biochemical items totaled 25 in all (see Table 3). Most of the parameter settings are software pre-settings, and the confidence factor for trimming changed from 0.25 to 0.3.

This study used 10-fold cross validation to validate the C4.5 decision tree model. Podgorelec et al. [20] also indicated that in medical data analysis, using sensibility and specificity to evaluate the prediction model for predicting classification accuracy is appropriate.

Fig. 7. Schematic diagram of temporal abstraction.
Sensitivity is the proportion of correctly predicted hospitalization of all patients, also known as the true positive fraction (TPF). Specificity is the proportion of samples correctly recognized as non-hospitalization of all patients without admission to hospital, also known as true negative fraction (TNF). Table 4 shows the formulae. The main purpose of this study is to find out the biochemical test value patterns causing patients’ hospitalization. Therefore, this study emphasizes sensibility evaluation.

4. Computational results

4.1. Parameters of TA

This research conducted an experiment using DT to compare two time intervals of TA including 3-month and 6-month intervals. For the latter, if a patient has a hospitalization record in July, then the experiment considered the biochemical value record from January to June. Experimental results show that the sensibility of a 6-month interval (31%) is much higher than the sensibility of a 3-month interval (13%). Furthermore, the performance of two different numbers of TA state were compared by also using DT, such as a 3-state (high, normal, low) TA versus a 7-state (extra high, high, higher than normal, normal, lower than normal, low, extra low) TA. The experimental results indicate that 7-state TA (81.60%) is better than that of a 3-state TA (71.50%).

4.2. Results of data mining

The number of inferred rules is 215 from the DT and 477 from the M5P via antecedent rules. After discussing and evaluating with five domain experts, twenty-six clinically meaningful rules listed in Tables 5 and 6 were filtered to better predict HD patients’ hospitalization. Some important rules are explained as follows.

4.2.1. Rules from the decision tree

Paths to each leaf of DT can be transformed into IF–THEN rules which are mutually exclusive and exhaustive on the IF parts. However, the rules are more complex than necessary. C4.5 employs a set of test samples independent from the training set to estimate the accuracy of each rule. A rule may be pruned by removing any condition in its antecedent that does not improve the estimate accuracy of the rule. For the following rules, the end of each rule in brackets includes the accuracy rate (\textit{accuracy}) for predicting patients’ hospitalization and the absolute number (\textit{occurrence}) of patients supporting the rule.

(1) Relevant rules to albumin

\textbf{IF} (Albumin = “normal(N) to extra low(XL)” \textbf{THEN} (Hospitalization = \textbf{YES}) (\textit{accuracy: 93%, occurrence: 12})

Albumin represents current nutrition status and inflammation degree and is one of the utmost factors influencing long-term survival. If albumin is less than 3.4 gm/dl, the patient will likely
Table 5
Clinically meaningful rules of HD patients’ hospitalization from DT.

<table>
<thead>
<tr>
<th>No</th>
<th>Rules</th>
<th>Accuracy</th>
<th>Occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Albumin = “lower than normal(N/L) to lower than normal(N/L)” AND Hbc = “normal(N) to low(L)”</td>
<td>100%</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>Albumin = “lower than normal(N/L) to low(L)” AND Triglyceride = “lower than normal(N/L) to low(L)” AND age &gt; 74</td>
<td>100%</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>Albumin = “normal(N) to extra low(XL)”</td>
<td>93%</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>Albumin = “low(L) to extra low(XL)”</td>
<td>92%</td>
<td>11</td>
</tr>
<tr>
<td>5</td>
<td>Albumin = “extra low(XL) to extra low(XL)”</td>
<td>83%</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>Albumin = “normal(N) to low(L)” AND Triglyceride = “normal(N) to low(L)” AND Platelet = “lower than normal(N/L) to normal(N)”</td>
<td>82%</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>Albumin = “lower than normal(N/L) to lower than normal(N/L)” AND Hbc = “normal(N) to lower than normal(N/L)”</td>
<td>78%</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>Albumin = “lower than normal(N/L) to low(L)” AND Triglyceride = “low(L) to low(L)” AND Kt/V = “normal(N) to normal(N)”</td>
<td>75%</td>
<td>12</td>
</tr>
<tr>
<td>9</td>
<td>Albumin = “normal(N) to lower than normal(N/L)”</td>
<td>71%</td>
<td>188</td>
</tr>
<tr>
<td>10</td>
<td>Albumin = “low(L) to low(L)”</td>
<td>71%</td>
<td>30</td>
</tr>
</tbody>
</table>

Table 6
Clinically meaningful rules of HD patients’ hospitalization from MSApriori algorithm.

<table>
<thead>
<tr>
<th>No</th>
<th>Rules</th>
<th>support</th>
<th>Conf.</th>
<th>occurrence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Diabetes = YES AND RBC = “lower than normal(N/L) to lower than normal(N/L)” AND URR = “high(H) to high(H)”</td>
<td>0.013</td>
<td>0.96</td>
<td>45</td>
</tr>
<tr>
<td>2</td>
<td>Diabetes = YES AND Ca × P = “lower than normal(N/L) to lower than normal(N/L)”</td>
<td>0.014</td>
<td>0.95</td>
<td>42</td>
</tr>
<tr>
<td>3</td>
<td>Diabetes = YES AND Ca × P = “lower than normal(N/L) to lower than normal(N/L)” AND P = “normal(N) to normal(N)”</td>
<td>0.02</td>
<td>0.93</td>
<td>70</td>
</tr>
<tr>
<td>4</td>
<td>Diabetes = YES AND Albumin = “normal(N) to lower than normal(N/L)” AND GlucoseAC = “higher than normal(N/H)”</td>
<td>0.012</td>
<td>0.92</td>
<td>44</td>
</tr>
<tr>
<td>5</td>
<td>Diabetes = YES AND MCV = “normal(N) to normal(N)” AND Albumin = “normal(N) to lower than normal(N/L)”</td>
<td>0.011</td>
<td>0.89</td>
<td>39</td>
</tr>
<tr>
<td>6</td>
<td>Albumin = “normal(N) to lower than normal(N/L)” AND Creatinine = “lower than normal(N/L) to lower than normal(N/L)”</td>
<td>0.012</td>
<td>0.89</td>
<td>42</td>
</tr>
<tr>
<td>7</td>
<td>RBC = “lower than normal(N/L) to low(L)” AND Albumin = “normal(N) to lower than normal(N/L)”</td>
<td>0.013</td>
<td>0.87</td>
<td>48</td>
</tr>
<tr>
<td>8</td>
<td>Hct = “normal(N) to lower than normal(N/L)” AND Albumin = “normal(N) to lower than normal(N/L)”</td>
<td>0.013</td>
<td>0.87</td>
<td>41</td>
</tr>
<tr>
<td>9</td>
<td>Diabetes = YES AND Hbc = “normal(N) to lower than normal(N/L)” AND Albumin = “normal(N) to lower than normal(N/L)”</td>
<td>0.012</td>
<td>0.87</td>
<td>45</td>
</tr>
<tr>
<td>10</td>
<td>Albumin = “normal(N) to lower than normal(N/L)”</td>
<td>0.065</td>
<td>0.83</td>
<td>188</td>
</tr>
<tr>
<td>11</td>
<td>Hbc = “normal(N) to lower than normal(N/L)” AND Hct = “normal(N) to lower than normal(N/L)”</td>
<td>0.035</td>
<td>0.81</td>
<td>100</td>
</tr>
<tr>
<td>12</td>
<td>RBC = “lower than normal(N/L) to low(L)” AND Hbc = “normal(N) to lower than normal(N/L)” AND Hct = “normal(N) to lower than normal(N/L)”</td>
<td>0.018</td>
<td>0.80</td>
<td>52</td>
</tr>
<tr>
<td>13</td>
<td>Hct = “normal(N) to lower than normal(N/L)”</td>
<td>0.079</td>
<td>0.77</td>
<td>213</td>
</tr>
<tr>
<td>14</td>
<td>Albumin = “lower than normal(N/L) to lower than normal(N/L)”</td>
<td>0.02</td>
<td>0.72</td>
<td>80</td>
</tr>
<tr>
<td>15</td>
<td>Diabetes = YES AND WBC = “normal(N) to higher than normal(N/H)”</td>
<td>0.018</td>
<td>0.70</td>
<td>66</td>
</tr>
</tbody>
</table>
If (Albumin = "normal(N) to lower than normal(N/L)")

THEN (Hospitalization = YES) (support:0.065; Conf:0.83, occurrence:188)

The same result obtained from the decision tree, too low albumin is still an important index for prognosis.

4.3. Summary

Although rules obtained from both DT and MSapriori algorithms are meaningful, we suggest using the MSapriori algorithm to find HD patients' hospitalization rules because of the following reasons: (1) The MSapriori algorithm allows missing value, whereas the DT algorithm does not allow any null values. (2) The MSapriori algorithm concentrates on one class, disregarding what happens to the other classes, whereas the DT split takes all classes into account, trying to maximize the purity of the split. (3) The set of extracted rules from DT is not much simpler than the corresponding DT. Therefore, we need to do some more work by pruning the resulting rule set, whereas rules from the MSapriori algorithm are easier. (4) Rules selected by the DT within a class are according to their estimated accuracy. These will result in lower occurrences. Comparison between Tables 5 and 6, the number of occurrence from DT is less than the one from the MSapriori algorithm.

5. Conclusion

Medical data analysis includes data gathering, preprocessing, result evaluation, favorable interaction, and discussions with clinicians for correct analytic results [4,6,8,13]. This study used data mining techniques for extracting professional knowledge. This method is an improvement over traditional face-to-face discussions with professional persons, and enables us to obtain important knowledge effectively and quickly.

The experimental results show that different data mining methods can be combined effectively, and more abundant patterns can be found for practical applications. Furthermore, we can add domain knowledge prior to data analysis by combining the TA method, to make mining results more likely comprehended by the clinician. Therefore, TA is an indispensable method for future medical time series data analyses.

Among the hospitalization patterns found in this study, albumin is the most important index for predicting patients' hospitalization. This index is also currently used clinically for predicting patients' death rate. The results of this study therefore have clinical significance. Predicting patients' hospitalization by biochemical value evolution of blood examination has been an undefined biochemical item in previous clinical applications. After validating by medical care personnel, the time evolution of this index value proves to have definite relevance for hospitalization.

This study combined TA with data mining techniques to analyze dialysis patients' biochemical data. The mined temporal hospitalization patterns are helpful for doctors to diagnose patient hospitalization probability and to suggest some immediate treatments to avoid hospitalization. Finally, we hope this research will help hemodialysis centers to improve health care quality.

Many relevant methods and concepts could be added for analysis results. Directions for future research include the following three points. First, Adlassnig et al. [1] indicated that combining TA with Fuzzy Logic is more coincident for describing actual temporal data situations and is also a direction worthy of study. Second, HD patients' hospitalization rules can be obtained either by one-class classification method [15] or by subgroup discovery algorithm [18]. Third, for mined temporal patterns, implementing a system to assist medical care personnel to carry out medical intervention or treatment, namely to automate the overall flow from temporal mining preprocessing to rule generating, would assist medical personnel with daily business.

Acknowledgement

We are grateful to the National Science Council for the grants program (NSC97-2221-E-415-008-MY3), and to the Hemodialysis Center of Chiayi Yang Ming Hospital for their professional consultations.

References

[19] B. Liu, W. Hsu, Y. Ma, Mining association rules with multiple minimum supports, Proc. of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-99), San Diego, CA, USA, 1999.